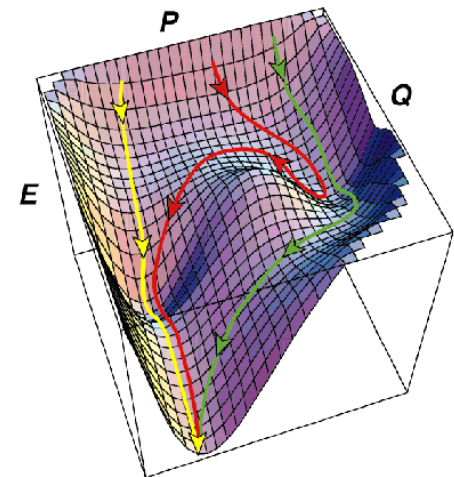


Homology Modelling

Revisited

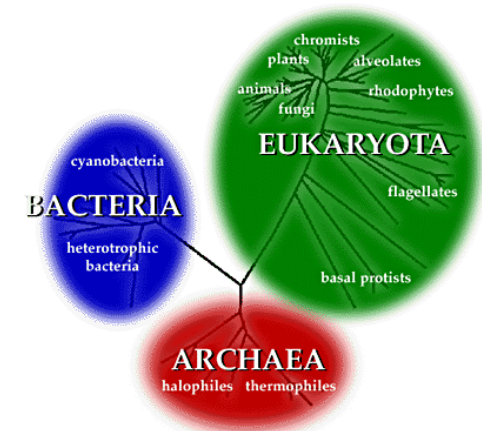
Why Do We Need Homology Modelling?

- *Ab Initio* protein folding (random sampling):
 - 100 aa, 3 conf./residue gives approximately 10^{48} different overall conformations!
- Random sampling is *NOT feasible*, even if conformations can be sampled at picosecond (10^{-12} sec) rates.
 - Levinthal's paradox
- Do homology modelling instead.



How Is It Possible?

- The structure of a protein is uniquely determined by its amino acid sequence
(but sequence is sometimes not enough):
 - prions
 - pH, ions, cofactors, chaperones
- Structure is conserved much longer than sequence in evolution.
 - Structure > Function > Sequence



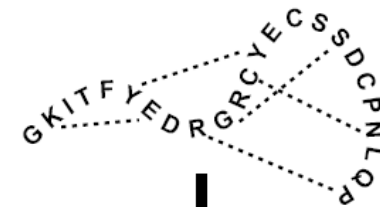
How Is It Done?

- Identify template(s)
 - Initial alignment
- Improve alignment
- Backbone generation
- Loop modelling
- Side chains
- Refinement
- Validation ←

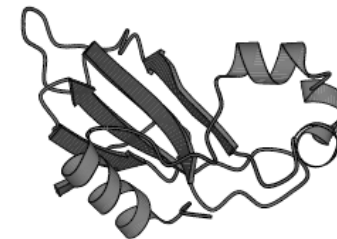
1. ALIGN SEQUENCE
WITH STRUCTURES:

3D GRISFFEDAGF-GHCYECSSDC-NLQP
3D GKITFYEDRGFGHCYECSSDC-NLQP
SEQ GKITFYEDRG---RCYECSSDCPNLQP

2. EXTRACT SPATIAL
RESTRAINTS:

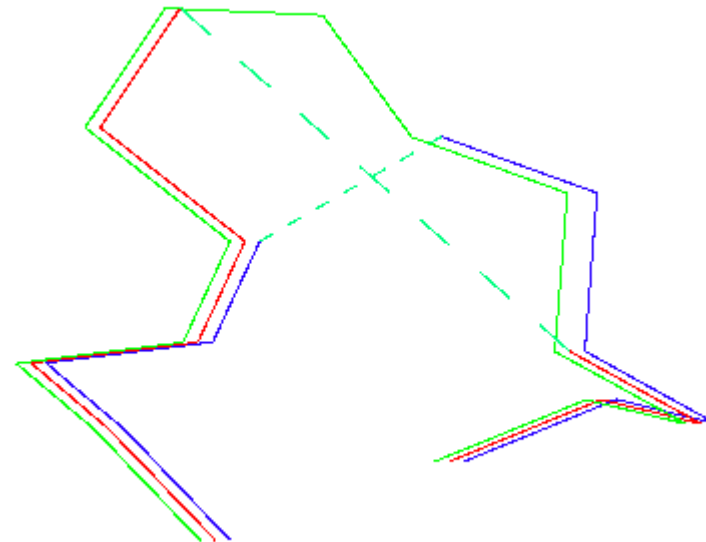


3. SATISFY SPATIAL
RESTRAINTS:



Improving the Alignment

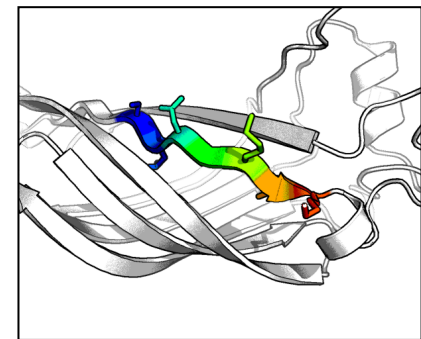
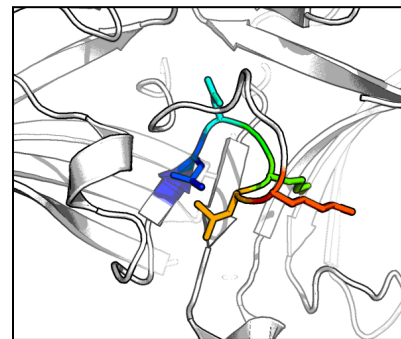
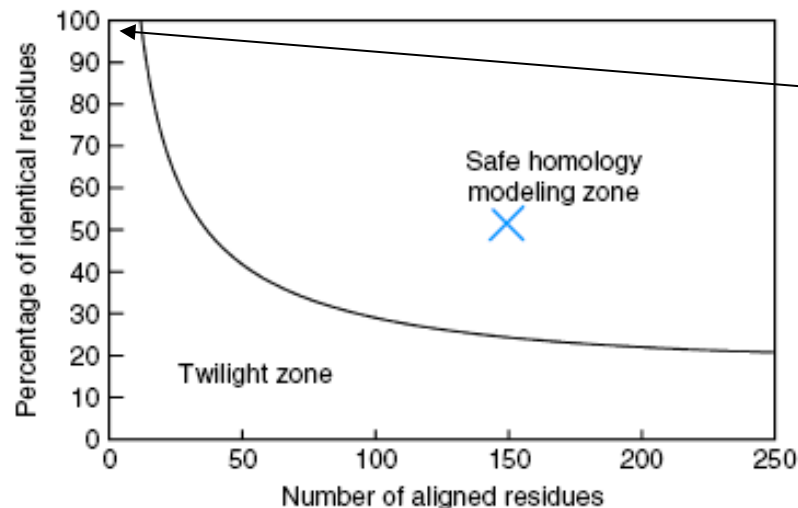
1	2	3	4	5	6	7	8	9	10	11	12	13	14
PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS
PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS
PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS



From "Professional Gambling" by Gert Vriend
<http://www.cmbi.kun.nl/gv/articles/text/gambling.html>

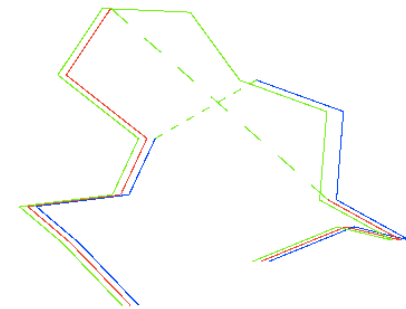
Template Quality

- Selecting the best template is crucial!
- The best template may not be the one with the highest % id (best p-value...)
 - Template 1: 93% id, 3.5 Å resolution ☹️
 - Template 2: 90% id, 1.5 Å resolution 😊



Error Recovery

- Errors in the model can NOT be recovered at a later step
 - The alignment can not make up for a bad choice of template.
 - Loop modeling can not make up for a poor alignment.
- The step where the errors were introduced should be redone.

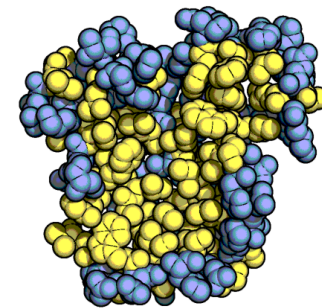
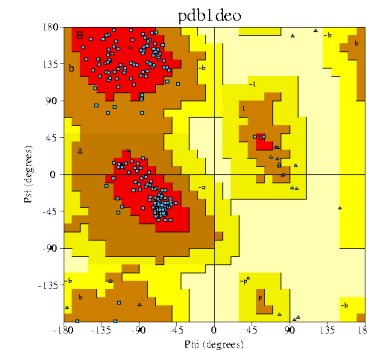
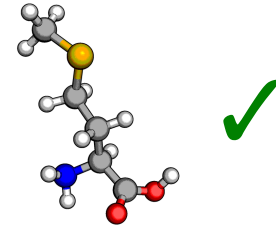


Thumb rules

- 1) no deletion with terminals far apart in space
- 2) no insertion in the core
- 3) check for conserved residues, especially G, P, C
- 4) check for motif conservation
- 5) be careful with secondary structure prediction
- 6) identify disordered/accessory regions
- 7) Be conservative with the first alignment

Validation

- Most programs will get the bond lengths and angles right.
- Model Rama. plot ~ template Rama. plot.
– select a high quality template!
- Inside/outside distributions of polar and apolar residues.



Model Validation – ProQ

- ProQ is a neural network-based predictor
 - Structural features → quality of a protein **model**.

- ProQ is optimized to find
 - correct models...
 - ...NOT (necessarily) native structures.

- Two quality measures:
 - MaxSub & LGscore

Arne Elofssons group: <http://www.sbc.su.se/~bjorn/ProQ/>

LGscore vs. MaxSub

- LG score

- Predict a structural alignment score:

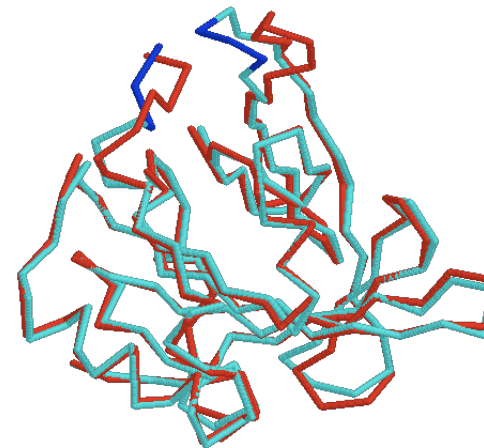
$$S_{str} = M \left(\sum \frac{1}{1 + (d_{i,j} / d_0)} - \frac{N_{gap}}{2} \right)$$

- Estimate the p-value (significance) of getting S_{str} compared to a random distribution.
- Report Lgscore as:

$$LGscore = -\log(p)$$

- MaxSub

- % of residues in model predicted to be within 3.5 Å of the true position in a superimposition with the true structure.



LGscore vs. MaxSub

Correct	Good	Very good
LGscore > 1.5	LGscore > 3	LGscore > 5
MaxSub > 0.1	MaxSub > 0.5	MaxSub > 0.8

Structure Validation Program Suites

- ProCheck

<http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/>

- WhatIf server

<http://swift.cmbi.ru.nl/whatif/>

Summary

- Successful homology modelling depends on the following:
 - Template quality
 - Alignment (add biological information)
 - Modelling program/procedure (use more than one)
- Always validate your final model!

Programme

8.00-8.10	Last week's Summary
8.10-8.20	Individual Quiz
8.25-8.50	Group Quiz
9.00-9.30	Fold recognition
9.30-9.40	Pause
9.40-12.00	Exercise

Fold Recognition and *Ab Initio* Protein Structure Prediction

*Based on slides by Pernille
Andersen*

- Threading and pair potentials
- *Ab initio* structure prediction methods
- **Human intervention** (what kind of knowledge can be used for alignment and selection of templates?)
- **Meta-servers** (the principle, 3d jury)
- Summary and take-home messages

Threading and Pair Potentials

- Compares a given sequence against known structures (folds)
- Potentials that describe tendencies observed in known protein structures

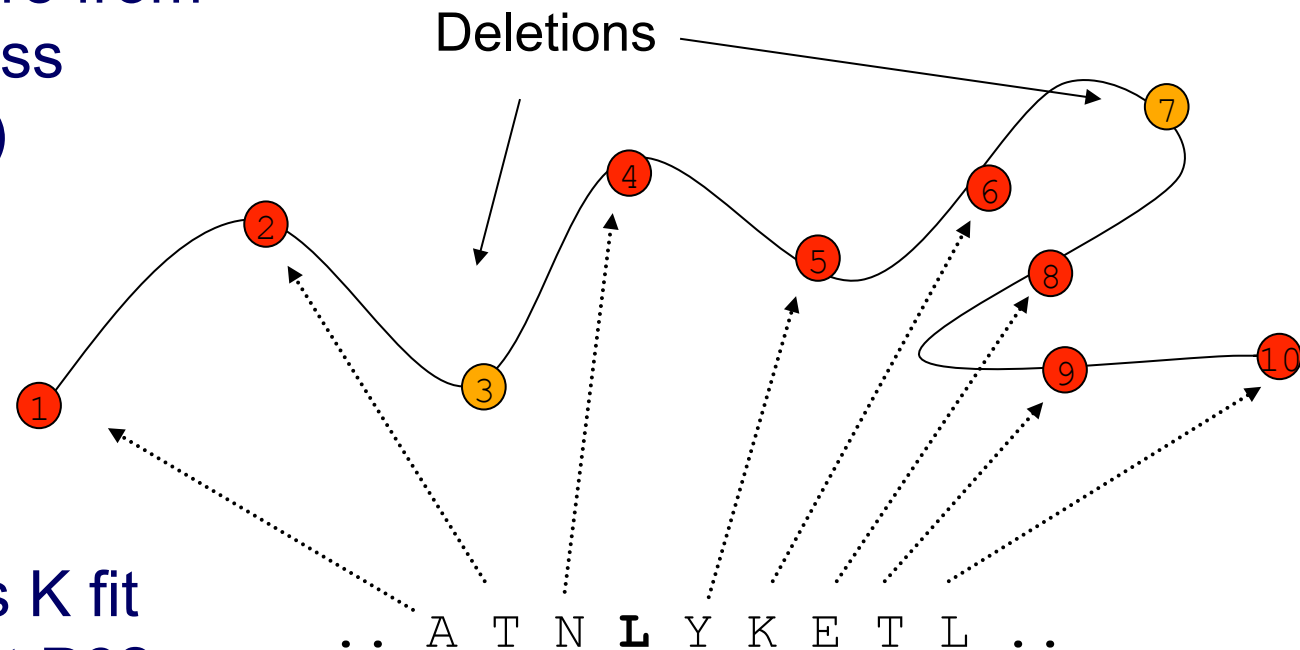
Example: Pair potentials
How normal is it to observe a pair of an alanine and a valine separated by 20 residues in the sequence and 3Å in space? (X)

How normal is it to observe any pair of residues separated by 20 residues and 3Å in space? (Y)

Potential: $E = -\log (X/Y)$

Potentials of Mean Force

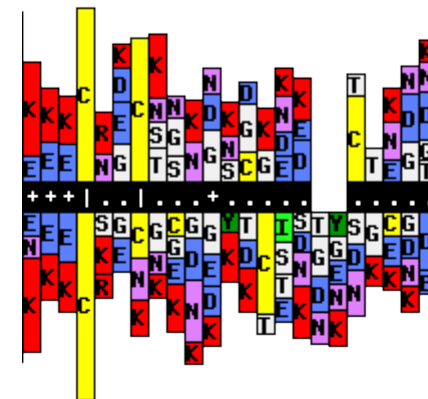
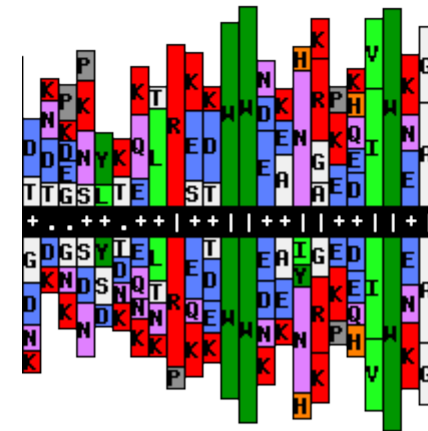
Alignment score from
structural fitness
(pair potential)



How well does K fit
environment at P6?
If P8 is acidic then
fine, if P8 is basic then
poor

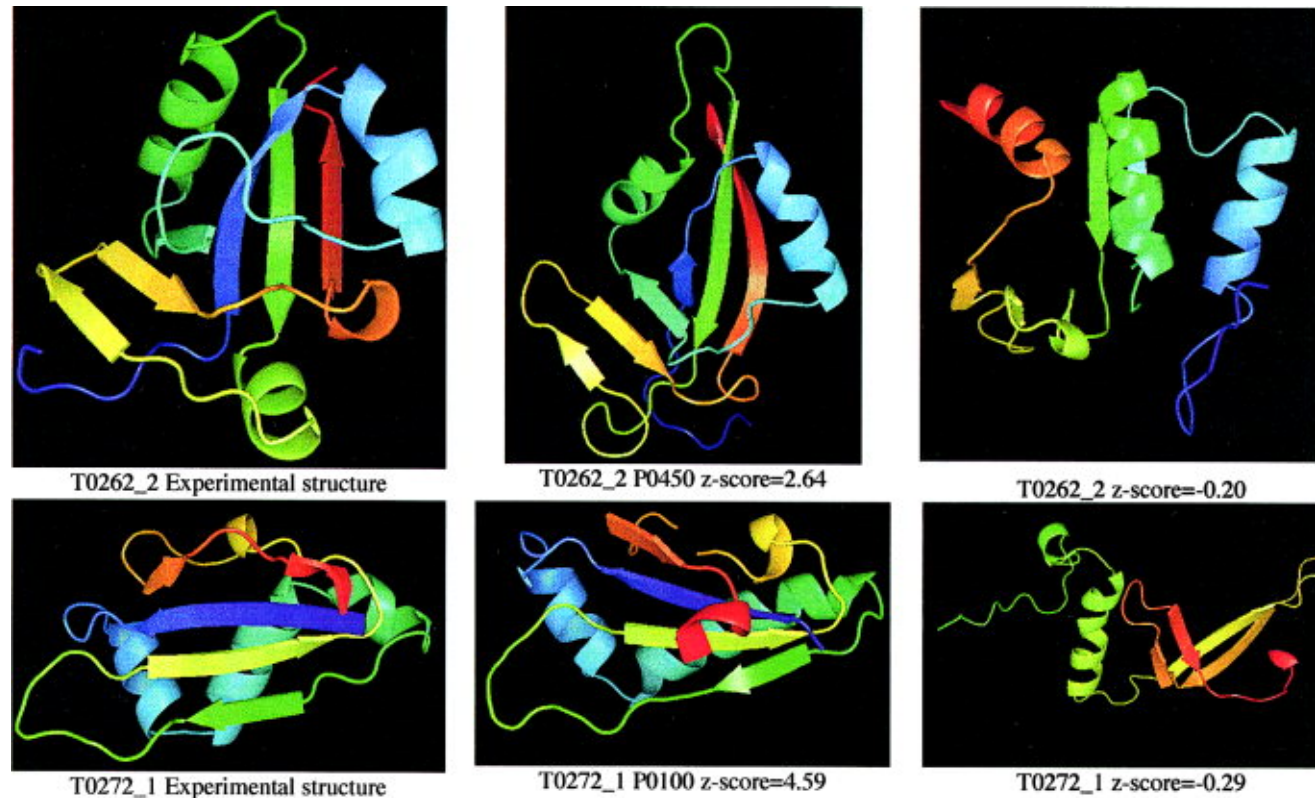
Threading Methods Today

- Problem: No protein is average
- Interactions in proteins cannot only be described by **pairs** of amino acids
- The information in the potentials is partly captured with sequence profiles or HMMs
- Today mostly used in **HYBRID** approaches in combination with profile-profile based methods
- Potentials can be used to score models based on different templates or alignments



HMM alignment,
hhpred

Fold Recognition Models in CASP



Two-high-scoring predictions by the top groups in FR/H (top) and FR/A (bottom). The assigned z-scores are given for the top predictions (center) as well as for two average predictions (right).

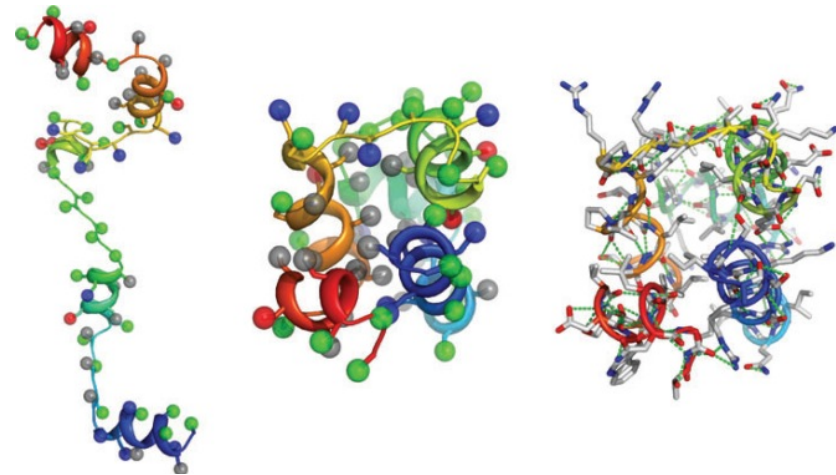
G. Wang Assessment of fold recognition predictions in CASP6, Proteins 61, S7, Pages 46-66

Ab Initio/ Free Modeling Methods

- Aim is to find the fold of native protein by simulating the biological process of protein folding.
- A VERY DIFFICULT task because a protein chain can fold into millions of different conformations.
- Use it **only** when no detectable homologues can be found.
- Methods can also be useful for fold recognition in cases of extremely low homology (e.g. convergent evolution).

Fragment-Based *Ab Initio* Modelling

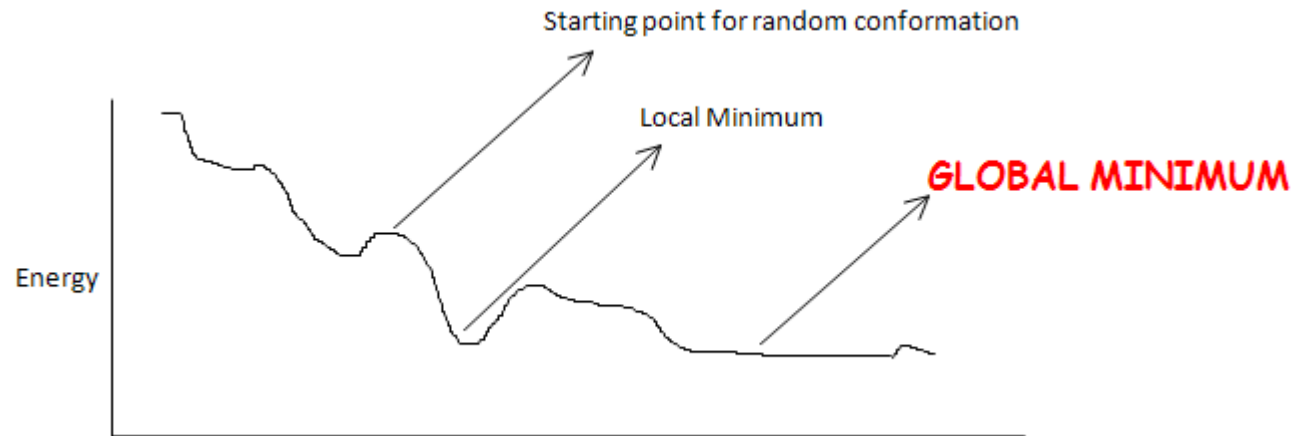
- Rosetta method of the Baker group:
 - Secondary structure prediction
 - Fragments library of 3 and 9 residues from known structures
 - Link fragments together, use only backbone and CB atoms
 - Contact/pair potential
 - Energy minimization techniques (Monte Carlo optimization) to calculate tertiary structure
 - Refine structure including side chains



Das R, Baker D, Annu. Rev. Biochem. 2008. 77:363–82

<http://robetta.bakerlab.org/>

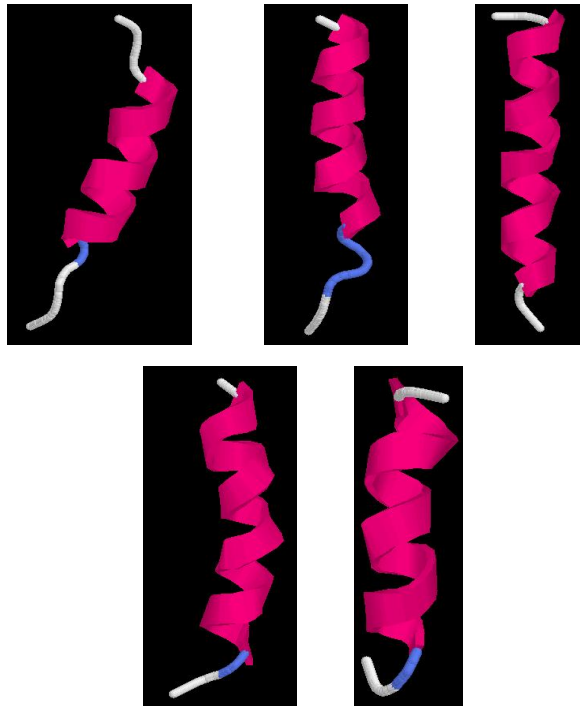
Energy Minimization



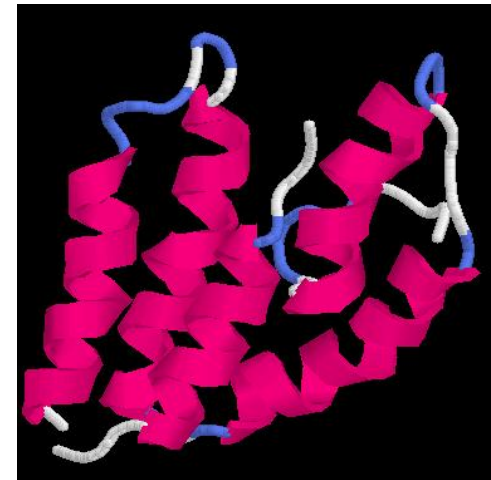
The energy of the whole protein model is minimized to obtain the final model

Problems with Empirical Potentials

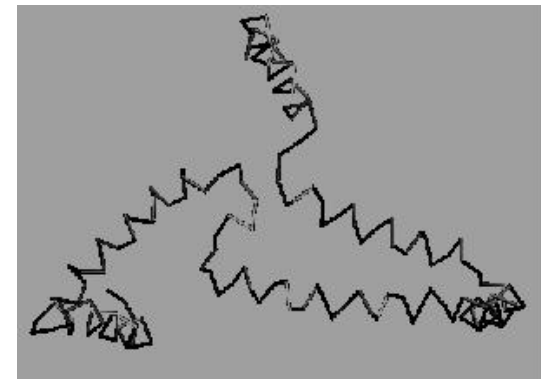
Fragments with
correct local structure



Nature's potential

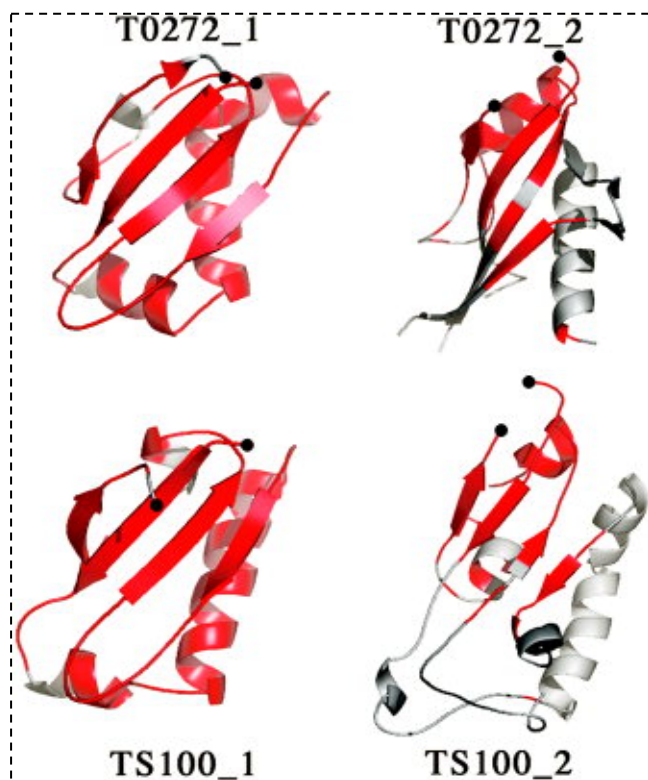


Empirical potential



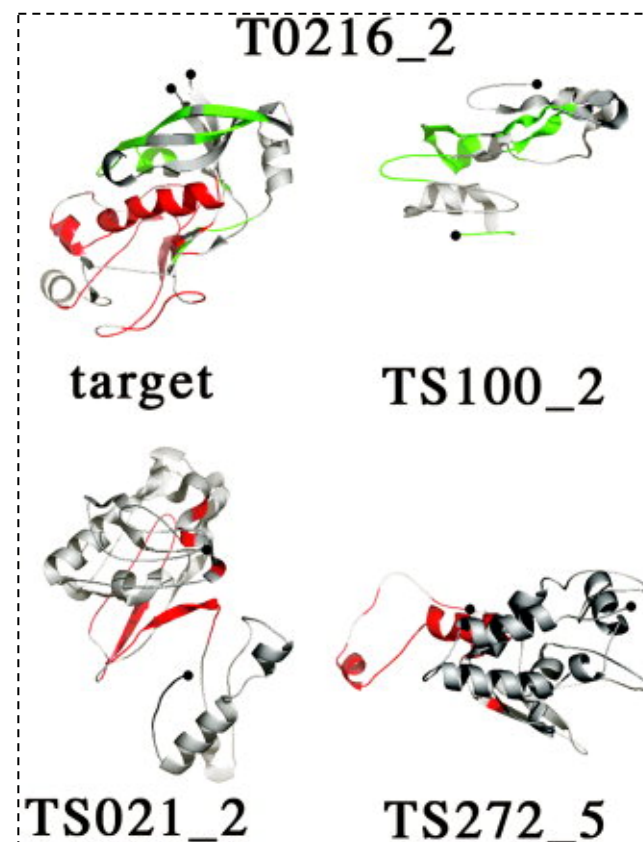
CASP6 & *Ab Initio* (new folds category)

Excellent modelling



The Baker group (#100) was among the top scoring

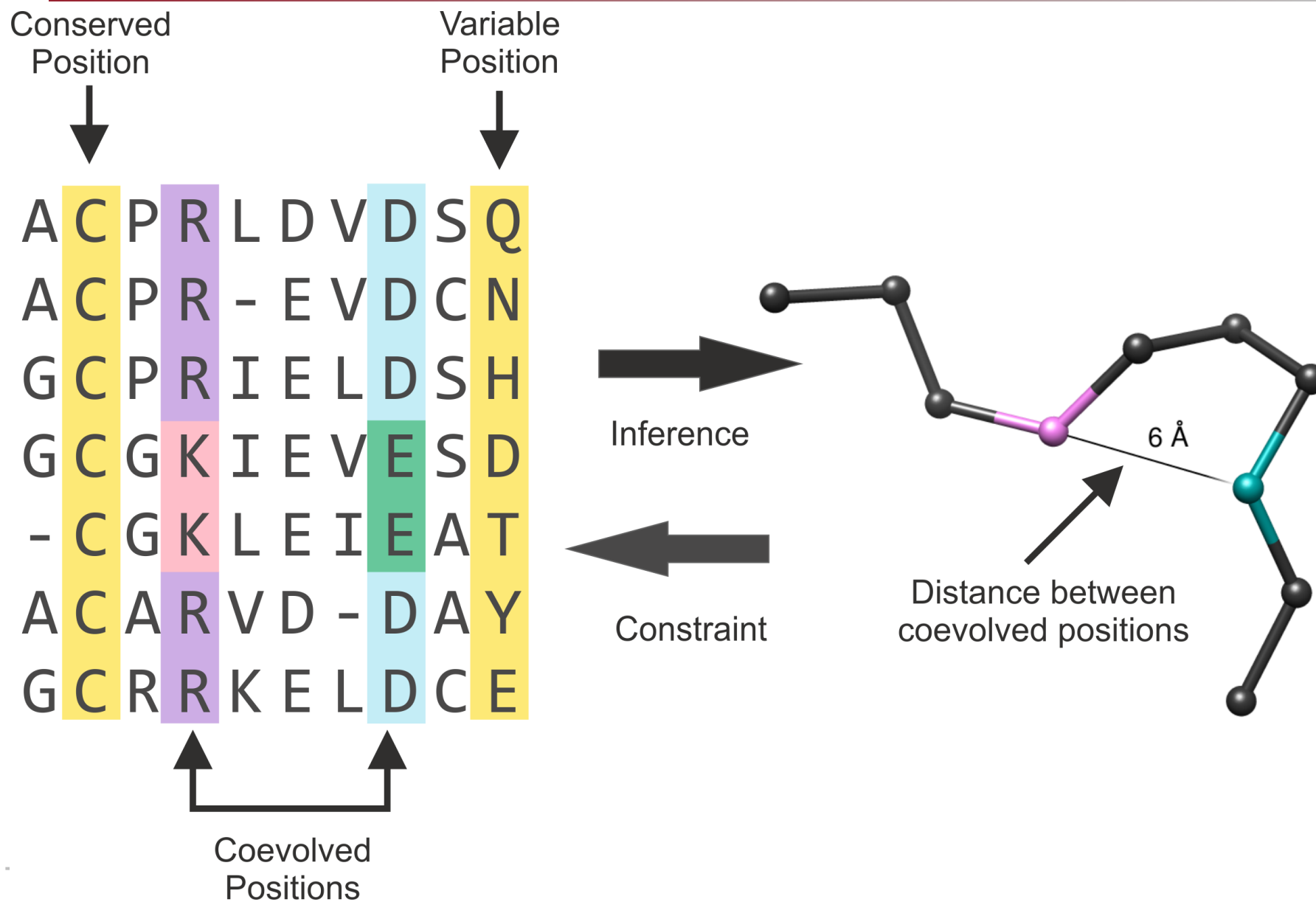
Hardest target



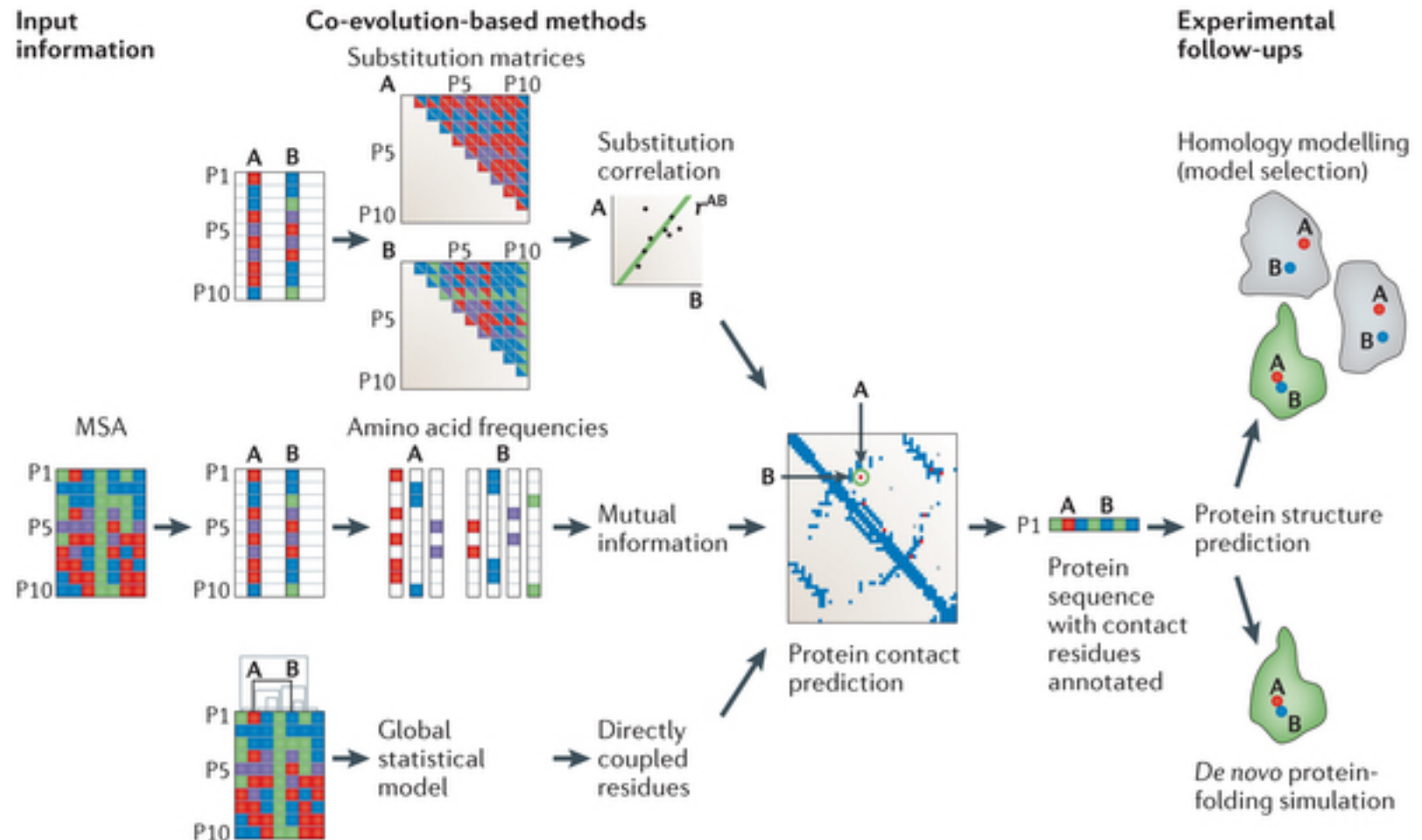
From coevolution



To function



Direct Coupling Analysis



Human Intervention

- The best groups in CASP use maximum knowledge of query proteins
- Specialists can help to find a correct template and correct alignments

Knowledge of function

Cysteines forming disulfide bridges or binding e.g. zinc molecules

Proteolytic cleavage sites

Other metal binding residues

Antibody epitopes or escape mutants

Ligand binding

Results from CD or fluorescence experiments

Meta-Servers

- Democratic modeling
 - The highest scoring hit is often wrong
 - Many prediction methods have the correct fold among the top 10-20 hits
 - If many different prediction methods all have the same fold among the top hits, this fold is probably correct

Server 1

Template 1 -> Model 1

Template 2 -> Model 2

Template 3 -> **Model 3**

Server 2

Template 1 -> **Model 1**

Template 2 -> Model 2

Template 3 -> Model 3

Server 3

Template 2 -> Model 1

Template 2 -> **Model 2**

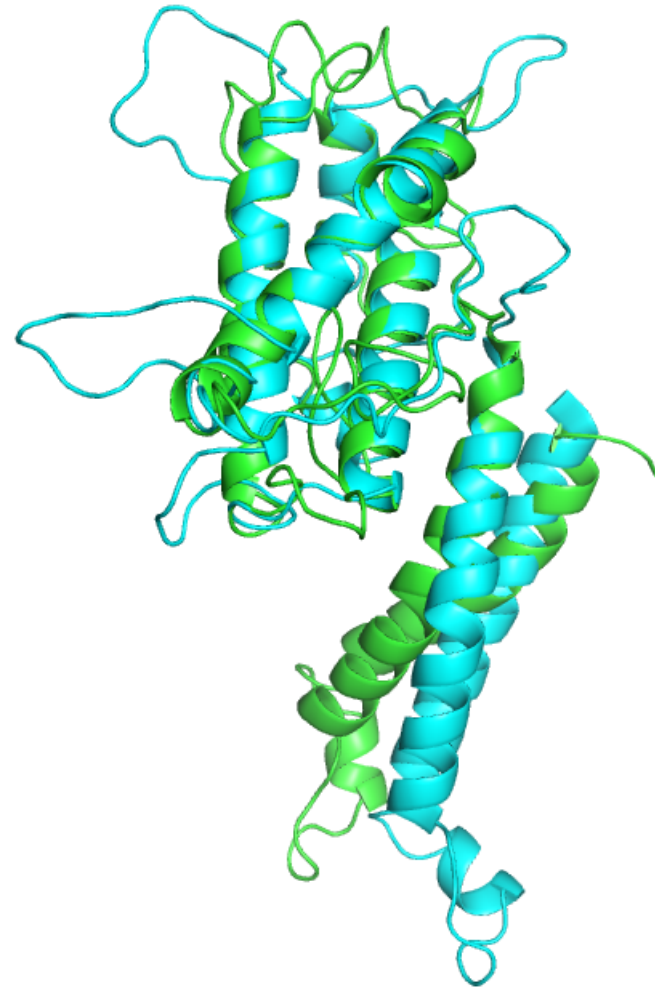
Template 3 -> Model 3

Example of a Meta-Server

- 3DJury http://meta.bioinfo.pl/submit_wizard.pl
 - Inspired by *Ab initio* modeling methods
 - Average of frequently obtained low energy structures is often closer to the native structure than the lowest energy structure
 - Find most abundant high scoring model in a list of prediction from several predictors
 1. Use output from a set of servers
 2. Superimpose all pairs of structures
 3. Similarity score based on # of C α pairs within 3.5Å
 - Similar methods developed by A. Elofsson (Pcons <http://pcons.net/>) and D. Fischer (3D shotgun)

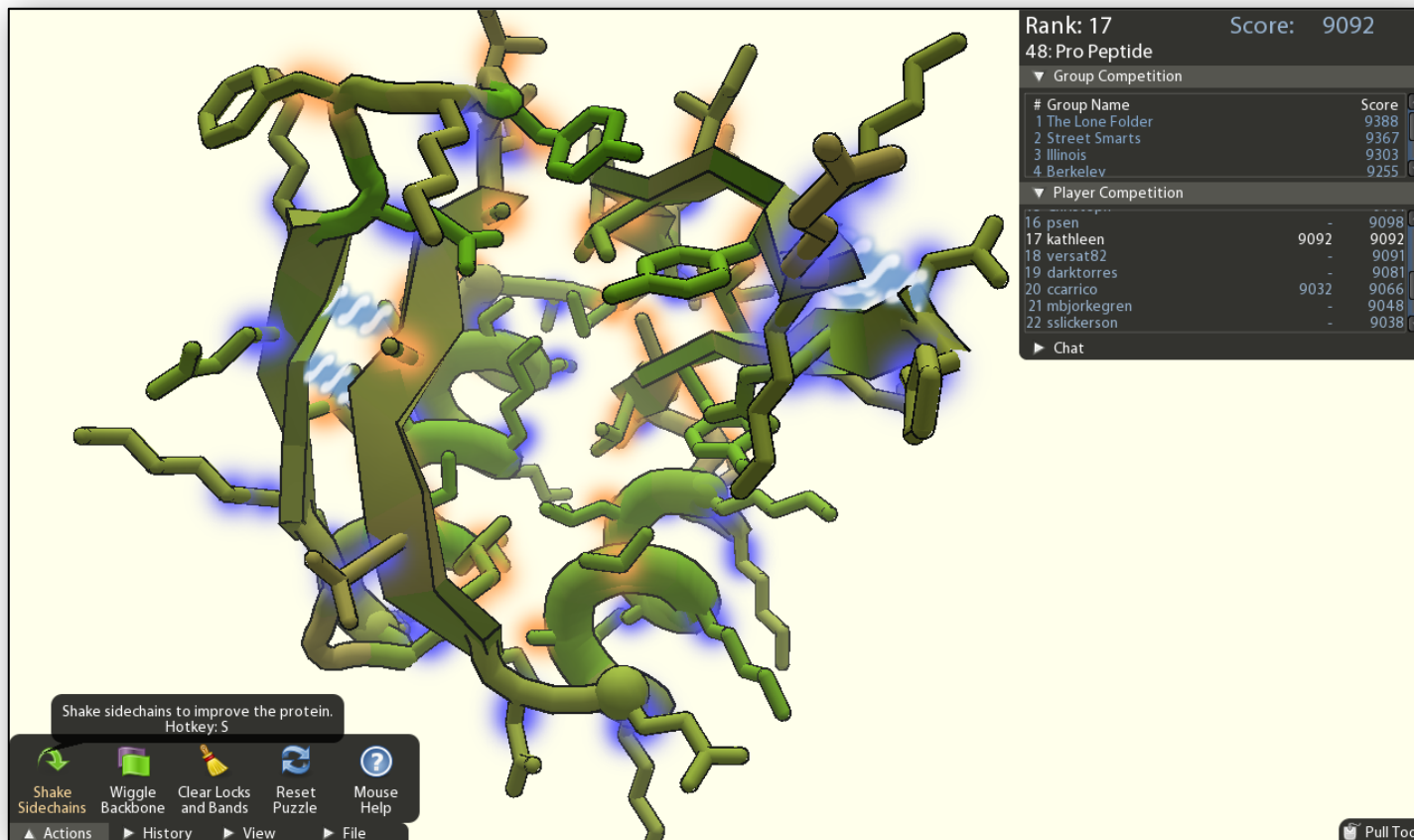
3DJury

- Because it is a meta-server it can be slow
- If queue is too long some servers are skipped
- Alternative conformations for a sequence are easily obtained



Human Intervention II

- **Fold It: The Protein Folding Game**



<http://fold.it/portal/>

Human Brain Power!

- Uses the HUMAN brain's pattern recognition resources for finding the lowest energy fold
- Can humans really help computers fold proteins?
 - *We're collecting data to find out if humans' pattern-recognition and puzzle-solving abilities make them more efficient than existing computer programs at pattern-folding tasks. If this turns out to be true, we can then teach human strategies to computers and fold proteins faster than ever!*

More FoldIt

- <http://www.youtube.com/user/uwfoldit>
- <http://fold.it/portal/>

Take Home Messages

- Hybrid methods using both threading methods and profile-profile alignments are the best
- Use only *Ab initio* methods if necessary and know that the quality is really low!
- Try to use as much knowledge as possible for alignment and template selections in difficult cases
- Use meta-servers when you can
- TRY FOLDIT!

Programme

8.00-8.10	Last week's Summary
8.10-8.20	Individual Quiz
8.25-8.50	Group Quiz
9.00-9.30	Fold recognition
9.30-9.40	Pause
9.40-12.00	Exercise

It came from the stable



It came from the stable



Your data

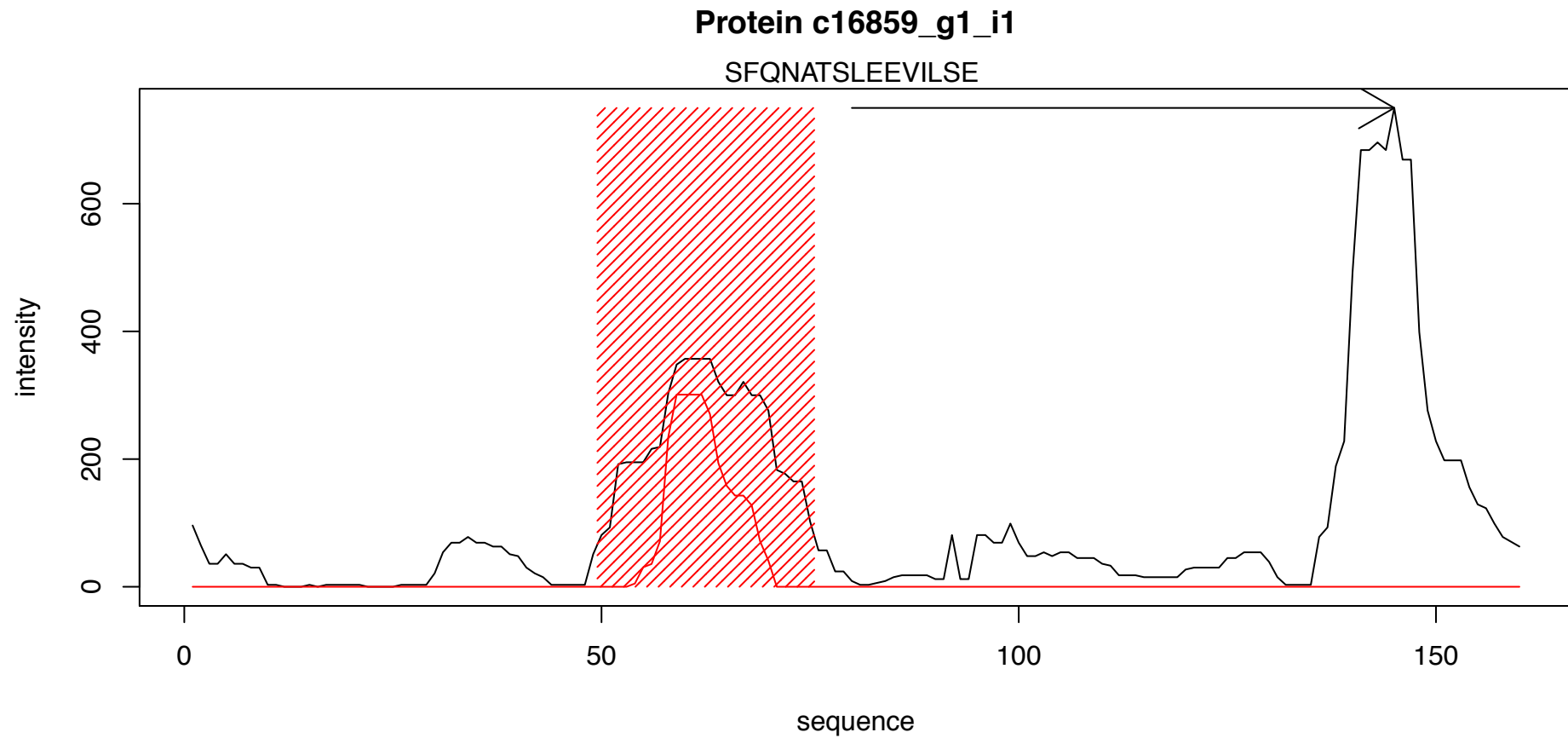
Metatranscriptomic

- bacterial community
- short/incomplete transcripts

Peptide chip

- linear antigenic peptides
- many false positives
- sticky peptides

Your data



Your goal

- make a preliminary study on 3/4 transcript
- choose the most promising one
- build a model and map the best epitope
- functionally characterize it

Your goal

- make a preliminary study on 3/4 transcript
- choose the most promising one
- build a model using different pipelines
(see wiki)
- map the best epitope
- functionally characterize it
(Pfam, TMHMM, SignalP, COG, Sifter)